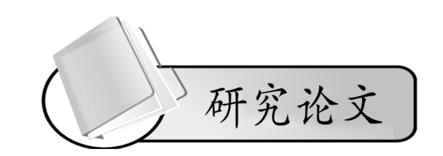


基于多知识库的短文本实体链接方法研究*
——以 Wikipedia 和 Freebase 为例

周鹏程[1]　武　川[1]　陆　伟[1,2]
[1](武汉大学信息管理学院　武汉　430072)
[2](武汉大学信息资源研究中心　武汉　430072)

摘要: 【目的】基于多知识库进行实体链接, 解决基于单一知识库的实体链接覆盖度低的问题。【方法】首先生成文本的 n-gram 并利用词性和多个指称–实体字典获取候选指称, 然后生成指称组合并保留覆盖度最大且不被其他组合包含的指称组合, 接着生成候选实体序列并利用多知识库信息计算实体序列的相关度, 最后选择相关度最大的实体序列为最终结果。【结果】以 Wikipedia 和 Freebase 为例的实验结果表明, 基于 Wikipedia+Freebase 的实体链接准确率、召回率、F 值分别达到 71.81%、76.86%、74.25%。【局限】基于词性过滤 n-gram 缺乏理论依据, 数据集 FACC1 具有高准确率和低召回率的特点。【结论】利用多个知识库的实体信息, 能够提升实体链接效果。

关键词: 实体链接　知识库　Wikipedia　Freebase
分类号: G353.1

1　引　言

实体(Entity)是现实世界中客观存在的并可以相互区别的事物, 既包括人名、地名、机构名等具体事物, 又包括概念、关系等抽象事物。实体链接(Entity Linking)是指文档中代表实体的文本片段, 即实体指称(Entity Mention, 简称指称), 与特定知识库(Knowledge Base)中的条目(Entry)相链接的过程, 有时称命名实体链接(Named Entity Linking)[1]。

实体广泛存在于各类文本中, 而面对未知实体时, 需要通过实体链接技术, 利用知识库中相关条目信息为原文本添加丰富的语义信息, 帮助读者加深关于该实体的了解, 从而有助于人或者计算机更好地理解、处理文本。

实体链接研究因其重要的研究意义而备受关注, 多项国际评测会议发布了实体链接相关的任务, 如 2007 年 INEX 会议发布的"Link the Wiki"任务(http://www.inex.otago.ac.nz/tracks/wiki-link/wiki-link.asp)、2009 年 TAC 会议发布的"Knowledge Base Population"任务(http://www.nist.gov/tac/)、2012 年 TREC 会议发布的"Knowledge Base Acceleration"任务(http://trec.nist.gov/)。实体链接在信息检索[2]、知识库构建[3]、问答系统[4]等领域都有较好的应用前景。

实体链接的难点在于两方面, 即多词一义和一词多义。多词一义是指实体可能有多个指称, 实体的标准名、别名、名称缩写等都可以用来指代该实体, 例如 Michael Jordan、MJ 和 Jordan 都可以指代实体 Michael Jeffrey Jordan。一词多义是指一个指称可以指

通讯作者: 周鹏程, ORCID: 0000-0002-5954-6863, Email: pc.zhou@whu.edu.cn。

*本文系国家自然科学基金面上项目"基于语言模型的通用实体检索建模及框架实现研究"(项目编号: 71173164)和武汉大学与中国科技信息研究所合作项目"科学文献的语义功能识别与深度利用"的研究成果之一。

代多个实体，例如 MJ 可能指代实体 Michael Jeffrey Jordan，也可以指代实体 Michael I. Jackson。解决一词多义问题要利用知识库中的实体信息进行实体消歧，单一知识库中的实体信息相对较少，笔者认为如果能利用多个知识库中的实体信息进行实体消歧，一词多义问题将会得到更好的解决。

知识库是实体链接研究的基础，常见的知识库包括 Wikipedia①、Freebase[5]、YAGO[6]、DBpedia[7]等，其中 Wikipedia 是实体链接研究中最常见的知识库，它包含丰富的文本语义信息，其中的每个实体页面都是某个实体的描述。Freebase 也比较常见，与 Wikipedia 相比，Freebase 中的实体信息更加结构化。谷歌于 2013 年发布了 Freebase 实体标注数据集 FACC1[8]，并且该数据集已经在信息检索领域[9]得到了应用。FACC1 是对 ClueWeb09②和 ClueWeb12③的实体标注，利用该数据集可以统计实体的流行度等信息。

本文提出了一种基于多知识库的实体链接方法，该方法利用多个指称–实体字典进行指称识别，利用多个知识库的实体信息进行实体消歧，以期解决基于单一知识库的实体链接覆盖度低的问题。

2 相关研究

实体链接包括两个步骤，即指称识别和实体消歧[10]。虽然有的研究[11]划分方式略有不同，但本质上是一样的。传统的实体链接大多关注长文档，近年来有研究者[9,12-14]开始关注短文本实体链接，如微博、查询词等，并已经在信息检索领域得到了应用[9]。二者的主要区别是短文本上下文信息少，实体消歧相对困难。另外短文本存在书写不规范问题，如丢失大小写信息和标点信息[9]、拼写错误[14]，这也给指称识别带来一定困难。因此笔者认为短文本实体链接研究更具挑战性，应给予更多的关注。

2.1 长文档实体链接

实体链接的第一步是进行指称识别，首先要构建一个指称–实体字典，大多数研究者抽取 Wikipedia 的实体页面、消歧页面、重定向页面的标题作为实体指称，建立指称–实体字典，还有其他的建立方式，如 Sil 等[15]抽取了 Freebase 中实体的标准名和别名。然后按一定的规则识别实体指称，如 Cucerzan[11]利用大小写规则、先验统计信息进行指称识别，并选择实体上下文与实体 Wikipedia 主页、候选实体之间的一致性最高的实体序列。Mihalcea 等[16]利用链接概率识别指称，然后综合利用知识工程方法和朴素贝叶斯分类方法确定最终的实体序列。

由于一个指称可能指向多个实体，因此需要用一定的方法确定指称所指向的实体，即实体消歧。目前实体消歧方法主要包括机器学习[17-18]、排序学习[19-21]、图模型[22-25]、无监督方法[11,26]和集成方法[27-28]等。Zhang 等[17]采用文献[11]的方法构建指称字典进行指称识别，如果候选指称集为空，则利用 Wikipedia 的“Did You Mean”和“Wikipedia Search Engine”特征补充候选指称集，并将实体消歧看作二类分类问题，即指称及其所指向的实体构成的指称–实体对为正例，与其他候选实体构成的指称–实体对为负例，选取词法特征、词–类别特征、实体类型等特征，采用 SVM 分类器分类，如果多个候选实体标记为正例，那么利用词袋、实体共现等特征计算指称–实体相似度，选择相似度最高的候选实体。Ratinov 等[20]假设指称已经给定，并提出两类特征：局部特征(即指称上下文与实体主页文本、指称所在文档与实体主页文本、指称上下文与实体上下文、指称所在文档与实体上下文等的余弦相似度)和全局特征(包括标准化谷歌距离、点互信息测度的实体类别相似度、入链相似度、出链相似度)，训练得到 Rank SVM 模型，选取排序最高的实体为该指称在上下文中所指的实体。Han 等[22]同样只关注实体消歧问题，以指称及其候选实体为节点，构建指称–实体、实体–实体关系图，利用类似 PageRank 的机制识别实体。

2.2 短文本实体链接

Ferragina 等[13]最早开始关注短文本实体链接，采

①https://www.wikipedia.org/.

②http://www.lemurproject.org/clueweb09.php/.

③http://www.lemurproject.org/clueweb12.php/.

用文献[11]的方法构建指称字典，并用人工规则过滤指称字典，利用该字典识别候选指称，然后利用指称指向实体的先验概率和候选实体与其他候选实体的相关性等特征，采用机器学习和人工规则两种方法进行实体消歧。Meij 等[14]则尽可能多地获得候选指称，提出 n-gram 特征、概念特征、n-gram–概念特征、Tweet 特征等 4 类特征，同样采用机器学习的方法识别概念并链向相应的 Wikipedia 页面。Liu 等[29]在 Meij 等的基础上又融合指称–指称特征，选择相似度得分最高的实体序列。

笔者发现目前实体链接研究都是基于单一知识库，但是由于某些实体只存在于特定的知识库库中，单一知识库可能无法完全覆盖文档中的实体。另外单一知识库可利用的实体信息相对较少，这将影响实体消歧的效果。针对以上问题，本文提出了一种基于多知识库的实体链接方法，该方法能够有效地利用多个知识库的实体信息，并同时对多个知识库进行实体链接。

3　实体链接

3.1　问题定义

实体链接是指给定一段文本，识别其中包含的指称，利用实体消歧方法确定指称指代的实体，并链向特定知识库中的相应条目。

实体链接问题的形式化定义如下：输入是由 n 个单词组成的文本 $t=(w_1,w_2\cdots w_n)$，输出是指称组合 $\vec{M}=(m_1,m_2\cdots m_l)$ 及其对应的实体序列 $\vec{E}=(e_1,e_2\cdots e_l)$，其中 e_i 表示特定知识库中的一个条目。如果 $|\vec{M}|=1$，那么输出是该指称可能对应的实体集合 $Set<e_1,e_2\cdots e_l>$。

3.2　实体链接方法

图 1 为基于多知识库的实体链接步骤，分为离线阶段和在线阶段，离线阶段构建指称–实体字典和实体映射字典；在线阶段为实体链接方法的主要步骤，包括生成 n-gram、候选指称识别、生成指称组合、生成实体序列、计算实体相关度等。

(1) 字典构建

为了进行指称识别，笔者从知识库中收集实体的标准名、别名等信息作为实体指称，进行相应的预处理，构建指称–实体字典。指称–实体字典包含两个域，

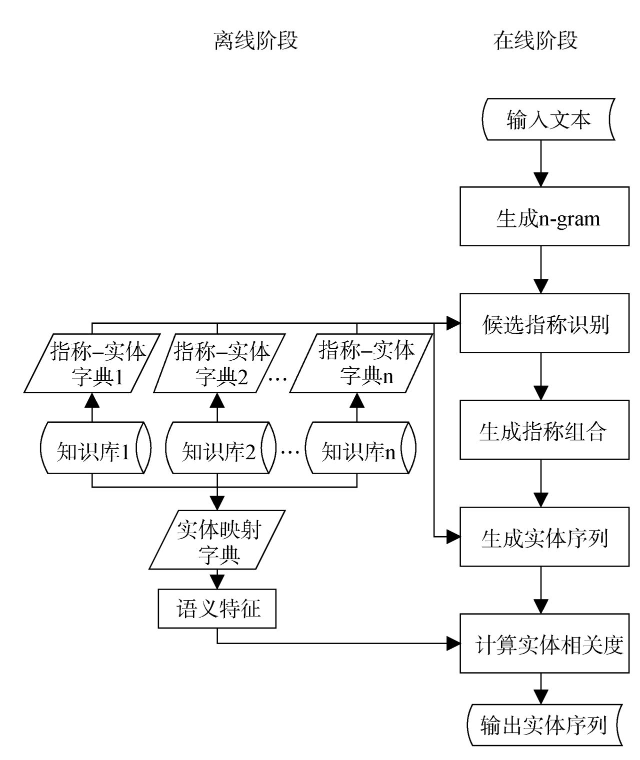


图 1　基于多知识库的实体链接

即指称域和实体域，存储格式为“$m\ \ e_1e_2\cdots e_n$”，其中 m 表示实体指称，$e_1e_2\cdots e_n$ 表示指称 m 可能指向的实体。

为了能同时利用多知识库信息进行实体链接，本文按照一定方法构建实体映射字典。实体映射字典包含 n 个域，存储格式为“$e_1e_2\cdots e_n$”，其中 e_i 表示知识库 i 的实体，且 $e_1e_2\cdots e_n$ 为不同知识库中的同一实体。

(2) 方法步骤

①生成 n-gram

生成输入短文本的 n-gram，例如对于短文本“obama family tree”，共生成 6 个 n-gram，即{obama, family, tree, obama family, family tree, obama family tree }。

②候选指称识别

对于生成的每个 n-gram，直接搜索多个指称–实体字典的指称域，如果任一指称–实体字典中存在相应的记录，则该 n-gram 可能是实体指称。如果 n-gram 所包含单词的词性都不是名词，那么将被过滤掉，因为根据笔者的观察，实体一般作为名词出现。例如对于 obama，指称–实体字典的指称域中存在相应的记录并且其词性为名词，因此 obama 是实体指称，同样，family、tree、obama family、family tree 也是实体指称。但是对于 obama family tree，指称–实体字典的指称域中不存在相应的记录，因此被过滤掉。从而“obama family tree”共保留 5 个可能的实体指称，即{obama, family, tree, obama family, family tree}。

③生成指称组合

候选指称识别阶段产生的实体指称可能存在重叠问题，有研究者[30]采用从左至右最长匹配的策略解决指称重叠问题，但是笔者认为这可能造成指称识别错误。本研究经过以下三个步骤生成候选指称组合：

1) 选择至少一个相互不重叠候选指称组成指称组合；

2) 保留覆盖度最大的指称组合；

3) 保留至少有一个指称不被其他组合包含的指称组合。这里的包含是指要么一个指称是另一个指称的一部分，要么两个指称相同。

例如"obama family tree"共保留两个指称组合，即{obama+family tree, obama family+tree}。

④生成实体序列及实体相关度计算

如果生成指称组合阶段共保留 n 个指称组合，且第 i 个指称组合包含 n_i 个指称，对于每一个指称，笔者合并各知识库中的候选实体记录，并取先验概率最大的 k 个实体为候选实体，则共生成 $\prod_{i=1}^{n} k^{n_i}$ 个实体序列。笔者认为同时出现的实体是相关的，因此计算每个实体序列的相关度得分并进行降序排列，返回得分最高的实体序列作为最终结果，见公式(1)：

$$\vec{e}^* = \arg\max_{\forall \vec{e} \in Set_{\vec{e}}} \alpha \cdot \sum \vec{\alpha} \cdot \vec{h}(m_i, e_i) + \beta \cdot \sum \vec{b} \cdot \vec{f}(\vec{e}) \qquad (1)$$

其中，$Set_{\vec{e}}$ 表示所有可能的实体序列；$\vec{h}(m_i,e_i)$ 表示指称 m_i 与其候选实体 e_i 的相关度函数，$\vec{a}$ 表示其权重向量，且有 $a_i \in (0,1)$ 及 $\sum a_i = 1$。$\vec{f}(\vec{e})$ 表示实体之间的相关度函数，$\vec{b}$ 表示其权重向量，且有 $b_i \in (0,1)$ 及 $\sum b_i = 1$。α、β 表示平衡两种相关度函数的权重，且有 α，$\beta \in (0,1)$ 及 $\alpha + \beta = 1$。

4 实例研究

由于在实体链接研究中 Wikipedia、Freebase 使用比较广泛，具有较强的代表性，因此笔者分别基于 Wikipedia、Freebase、Wikipedia+Freebase 进行实体链接。

4.1 基于 Wikipedia 的实体链接

(1) 指称–实体字典构建

本文采用 Bunescu 等[31]的方法，抽取 Wikipedia 的实体页面、消歧页面、重定向页面的页面标题以及实体主页的锚文本，进行小写化等预处理，构建指称–实体字典。同时，利用锚文本统计指称指向其候选实体的次数，并将其存入指称–实体字典中。

由于本文的输入不包含特殊符号并且都是小写化的，为了与输入进行匹配，笔者移除了实体指称中的特殊符号，并进行小写化处理，同时移除指称中的消歧信息，例如对于指称 The Last Supper (Leonardo da Vinci)，括号及其内部的信息被移除，这是因为文档提及实体时，通常不会附带实体的消歧信息。如果经过上述的处理后，两个指称变成了相同指称，那么合并指称记录，例如：

bilos Daniel_Ruben_Bilos_(2) Daniel_Rubén_Bilos_(1) Daniel_Bilos_(3)

其中 bilos 是指实体指称，存储于指称域；Daniel_Ruben_Bilos、Daniel_Rubén_Bilos、Daniel_Bilos 是 bilos 可能指向的实体(这里用实体的 Wikipedia 主页标题表示)，且在 Wikipedia 的主页文本中 bilos 指向三个实体的次数分别是 2 次、1 次、3 次，实体及次数信息存储于实体域，利用次数信息可计算候选实体的先验概率，例如 bilos 指向 Daniel_Bilos 的先验概率为：

$$Prior(Daniel|bilos) = \frac{3}{2+1+3} = 0.5 \qquad (2)$$

(2) Wikipedia 实体特征

①指称–实体特征

1) 先验概率

候选实体的先验概率是重要的消歧信息，很多研究[23,29,31]都有使用，笔者利用公式(2)计算候选实体的先验概率。注意，本文取各候选实体先验概率的算术平均数为实体序列的先验概率，字符串相似度的计算类似。

2) 字符串相似度

如果指称与其候选实体的标准名的相似度越高，那么指称指向该候选实体的概率越大。编辑距离(Edit Distance)是一种度量字符串相似度的方法，它是指一个字符串转变成另一个字符串所需要的最小编辑操作次数。对于两个给定的字符串，编辑距离越小表示两个字符串相似度越高。本文用公式(3)计算指称与其候选实体标准名的相似度。

$$h_{str}(m_i, CN(e_i)) = 1 - \frac{ED(m_i, CN(e_i))}{Max\{length(m_i), length(CN(e_i))\}} \qquad (3)$$

其中 $CN(e_i)$ 表示实体 e_i 的标准名，即 Wikipedia 实体主页标题；$h_{str}(m_i,CN(e_i))$ 表示指称与其候选实体标准名的字符串相似度，该值越高表示二者相似度越大；$ED(m_i,CN(e_i))$ 表示指称与其候选实体标准名的编辑距离，$Max\{length(m_i), length(CN(e_i))\}$ 表示指称与其候选实体标准名字符串长度较大者。

②实体–实体特征

1) 文本相关度

如果两个实体相关，那么它们的实体描述文本可能会讨论相同的内容，因此文本相关度可以用来表征实体相关

度。笔者对 Wikipedia 实体主页文本做了小写化、移除特殊字符、去除停用词等处理，用公式(4)计算两个实体之间的文本相关度。

$$f_t(e_i,e_j)=\frac{\sum_{k=1}^{l} w_{ik}\cdot w_{jk}}{\sqrt{\sum_{k=1}^{l} w_{ik}^2}\cdot\sqrt{\sum_{k=1}^{l} w_{jk}^2}} \tag{4}$$

其中 $f_t(e_i,e_j)$ 表示实体 e_i 与 e_j 的文本相关度，l 表示两段文本单词总数，w_{ik} 表示第 k 个单词在第 i 篇文档中的权重，本文将单词的频率作为其权重。注意，这里对实体文本相关度的定义只考虑两个实体的情况，如果实体序列包含的实体数大于 2，那么取两两实体的文本相关度的算术平均数为实体序列的文本相关度，其他相关度计算与之类似。

2) 相关实体相关度

如果两个实体相关，那么它们可能会存在相同的相关实体，因此相关实体相关度可以用来表征实体相关度。Wikipedia 的实体主页中存在指向其他实体页面的链接，可以利用这些链接搜集候选实体的相关实体集。

Wikipedia 实体 e_i 存在三种类型的相关实体：

• 入链相关实体，即实体 e_i 在实体 e_j 的主页中出现，而实体 e_j 在实体 e_i 的主页中未出现，则实体 e_j 是实体 e_i 的入链相关实体。

• 出链相关实体，即实体 e_j 在实体 e_i 的主页中出现，而实体 e_i 在实体 e_j 的主页中未出现，则实体 e_j 是实体 e_i 的出链相关实体。

• 互指相关实体，即实体 e_j 在实体 e_i 的主页中出现，且实体 e_i 在实体 e_j 的主页中也出现，则实体 e_j 是实体 e_i 的互指相关实体。

本文利用 Jaccard 系数表示两个实体的相关实体相关度，公式如下：

$$f_{reo}(e_i,e_j)=\frac{|Set_{oi}\cap Set_{oj}|}{|Set_{oi}\cup Set_{oj}|} \tag{5}$$

其中 $f_{reo}(e_i,e_j)$ 表示实体 e_i 和实体 e_j 的出链相关实体相关度，Set_{oi} 和 Set_{oj} 分别表示实体 e_i 和实体 e_j 的出链相关实体集合。公式(5)以出链相关实体为列，入链相关实体相关度、互指相关实体相关度的计算公式与之类似，实体 e_i 和实体 e_j 的相关实体相关度由三种类型相关实体相关度加权平均得到。

3) 类别相关度

如果两个实体相关，那么它们可能属于同一类别，因此类别相关度可以用来表征实体相关度。Wikipedia 的编辑者为每个实体标注了若干类别，类别信息可以从实体主页中获取。

仍利用 Jaccard 系数表示两个实体的类别相关度，公式如下：

$$f_c(e_i,e_j)=\frac{|Set_{ci}\cap Set_{cj}|}{|Set_{ci}\cup Set_{cj}|} \tag{6}$$

其中 $f_c(e_i,e_j)$ 表示实体 e_i 和实体 e_j 的类别相关度，Set_{ci} 和 Set_{cj} 分别表示实体 e_i 实体 e_j 的类别集合。

4.2 基于 Freebase 的实体链接

(1) 指称–实体字典构建

本文抽取 Freebase 实体的标准名和别名构建指称–实体字典。Freebase 的实体信息结构化程度高，有专门存储实体属性的字段，可以直接从名称字段和别名字段抽取实体标准名和别名。同样，笔者对 Freebase 实体指称做了小写化和去除特殊符号的处理，并利用 ClueWeb09 的 Freebase 实体标注数据集 FACC1[8]计算指称指向其候选实体的次数，例如：

baldwin_vi /m/0129jf_(3) /m/01_dt9_(6)

其中 baldwin_vi 是指实体指称，存储于指称域；/m/0129jf、/m/01_dt9 是指 baldwin_vi 可能指向的实体(这里用 Freebase 唯一标识符表示)，且在 ClueWeb09 中 baldwin_vi 指向两个实体的次数分别是 3 次、6 次，实体及次数信息存储于实体域，实体先验概率可用公式(2)计算。

(2) Freebase 实体特征

①指称–实体特征

这里仍采用先验概率和字符串相似度两个特征，定义及含义均同 Wikipedia 实体。注意，Freebase 实体标准名是从名称字段中抽取。

②实体–实体特征

1) 文本相关度

Freebase 实体文本相关度的文本取自实体描述(Description)字段，其含义、预处理、计算公式均同 Wikipedia 实体。

2)类型相关度

类似于 Wikipedia 实体的类别，如果两个实体相关，那么它们可能属于同一类型，因此类型相关度可以用来表征实体之间的相关度。Freebase 的编辑者为每个实体标注了若干类型，类型信息存储在类型字段中。例如实体 Barack Obama (Freebase 唯一标识符是/m/02mjmr)被标注了/people/person、/government/politician、/award/award_winner 等 97 种类型。可以看出，Freebase 实体可能有多种类型，并且每种类型是分层次的。笔者对实体的各层次类型名做了简单的词频统计，词频作为权重，仍用公式(4)计算两个实体的类型相关度。

4.3 基于 Wikipedia+Freebase 的实体链接

基于 Wikipedia+Freebase 的实体链接同时利用 4.1

节和 4.2 节中的指称–实体字典进行候选指称识别。由于 Wikipedia 与 Freebase 包含了实体不同方面的信息，因此本文从两个知识库中抽取了不同的实体特征，且二者可形成互补关系。为了能够同时利用 Wikipedia 和 Freebase 实体特征，本文构建了 Wikipedia 实体与 Freebase 实体映射字典。

(1) Wikipedia 实体与 Freebase 实体映射字典

Freebase 的实体页面中存在等价页面(Equivalent Webpage)域，其中包含了与之等价的其他知识库链接。笔者抽取与之等价的 Wikipedia 实体页面的标题，从而建立了 Wikipedia 实体与 Freebase 实体一一对应的关系，例如：

/m/03kkbz　873558　Ivan_Bella

其中/m/03kkbz 是 Freebase 实体唯一标识符，可通过该标识符获取 Freebase 实体的语义信息，如别名、类型、描述等；Ivan_Bella 是与/m/03kkbz 等价的 Wikipedia 实体页面标题，利用该标题可获取 Wikipedia 实体的语义信息，如类别、出/入链、主页文本等; 873558 是 Wikipedia 实体的编号。

(2) Wikipedia+Freebase 实体特征

对于指称–实体特征，仍采用先验概率和字符串相似度，这里对 Wikipedia 和 Freebase 相应的特征做算术平均；对于实体–实体特征，利用 Wikipedia 实体与 Freebase 实体映射字典，基于 Wikipedia+Freebase 的实体链接融合 Wikipedia 的文本、相关实体、类别等相关度和 Freebase 的类型相关度进行实体消歧，具体定义见 4.1 节和 4.2 节。

5　实验及结果

5.1　数据集与预处理

实验的输入是 2009 年–2012 年国际文本检索会议(TREC)的 Web Track 任务中的 200 个查询主题(Topic)，对其中包含的实体进行人工标注。由于部分查询主题本身即为实体指称，无法根据上下文进行实体消歧，因此将其移除，最终共处理 179 个查询主题，标注 242 个实体。

笔者下载 2013 年 12 月 2 日 Wikipedia 的转储文件①，包含实体页面、消歧页面、重定向页面等共 4 450 000 余篇，以及页面链接关系、实体类别信息等。利用 Java②语言处理原始文件、抽取实体指称、计算指称指向其候选实体的先验概率，用 Lucene③建立索引，以方便指称搜索，从 Wikipedia 中共抽取 14 870 000 多条指称。同时下载 2014 年 7 月 6 日 Freebase 的 RDF 文件④，包含实体的别名、等价页面、描述等属性，共包含 42 660 000 多个实体。利用同样的工具完成指称抽取、索引建立等工作，从 Freebase 中共抽取 22 130 000 多条指称。

由于 Freebase 中存在其与 Wikipedia 的映射关系，笔者抽取该映射关系，并利用 Lucence 构建相应的索引。

5.2　实验结果的评价指标

采用准确率、召回率和 F 值评价实验效果，三个指标的定义如下：

$$\text{Precision} = \frac{|Set_R \cap Set_L|}{|Set_R|} \tag{7}$$

$$\text{Recall} = \frac{|Set_R \cap Set_L|}{|Set_L|} \tag{8}$$

$$\text{F-value} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precsion} + \text{Recall}} \tag{9}$$

其中 Set_R 表示利用本文的方法识别的实体集合，Set_L 表示标注实体集合，$|Set_L|$表示集合中元素的个数，$|Set_R|$、$|Set_R \cap Set_L|$ 的含义类似。Precision 表示准确率，即正确识别的实体数占识别实体总数的比例；Recall 表示召回率，即正确识别的实体数占标注实体总数的比例，F-value 表示准确率和召回率的调和平均数。

5.3　实验结果

从表 1 可以看出，基于 Wikipedia 的实体链接，准确率达到 62.68%，召回率达到 71.49%，F 值达到

①https://dumps.wikimedia.org/.
②http://www.oracle.com/technetwork/java/index.html.
③http://lucene.apache.org/.
④https://developers.google.com/freebase/data.

66.80%；基于 Freebase 的实体链接，准确率、召回率、F 值分别达到 69.32%、75.62%、72.33%；基于 Wikipedia+Freebase 的实体链接，准确率、召回率、F 值分别达到 71.81%、76.86%、74.25%。

表 1 基于不同知识库以及同时基于两个知识库的实体链接评测结果

知识库	准确率	召回率	F 值
Wikipedia	62.68%	71.49%	66.80%
Freebase	69.32%	75.62%	72.33%
Wikipedia+Freebase	**71.81%** (+14.57%) (+3.59%)	**76.86%** (+7.51%) (+1.64%)	**74.25%** (+11.15%) (+2.65%)

（注：最好的结果加粗表示。括号内数值分别表示基于 Wikipedia+Freebase 的实体链接效果相对基于 Wikipedia 或 Freebase 的实体链接效果的提升值。）

5.4 讨 论

实验结果显示，基于 Wikipedia+Freebase 的实体链接效果均高于基于 Wikipedia 或 Freebase 的效果，其中准确率分别提升 14.57%和 3.59%，召回率分别提升 7.51%和 1.64%，F 值分别提升 11.15%和 2.65%，实验结果证明了基于多知识的实体链接方法的有效性。图 2 和图 3 显示了 15 个查询主题在三组实验中的召回率和准确率，可以看出，在基于 Wikipedia 的实体链接实验中，部分查询主题效果较好，例如 pacific northwest laboratory、arkadelphia health club；在基于 Freebase 的实体链接实验中，部分查询主题效果较好，例如 condos in florida、uss yorktown charleston sc；在基于 Wikipedia+Freebase 的实体链接实验中，15 个查询主题的效果都较好。

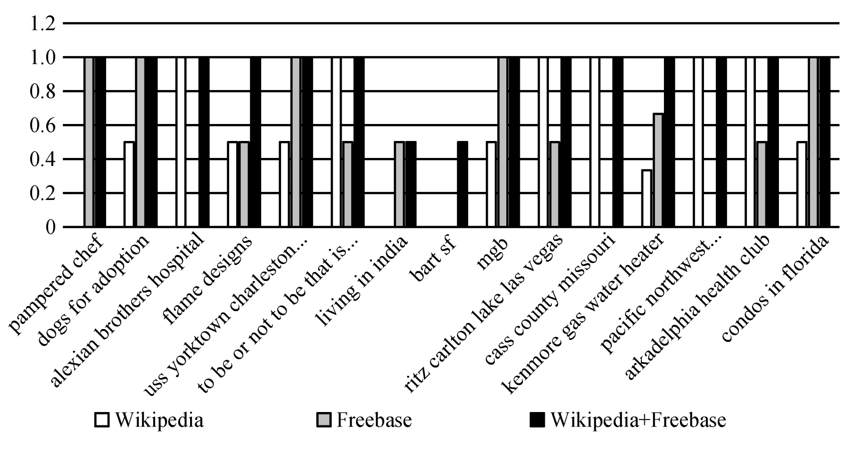


图 2 15 个查询主题在三组实验中的召回率

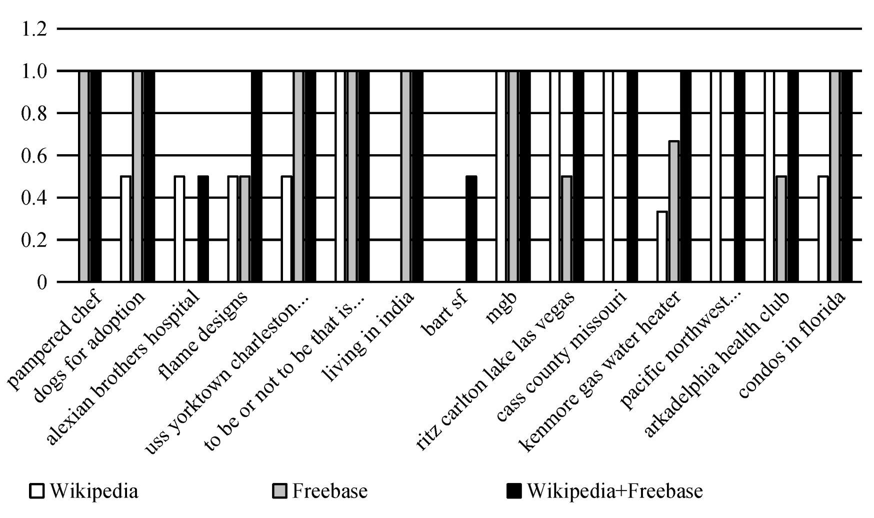


图 3 15 个查询主题在三组实验中的准确率

为了分析效果提升的原因，笔者补充了4组实验，结果如表2所示：

表2　基于不同知识库以及补充实验评测结果

知识库	准确率	召回率	F值
Wikipedia	62.68%	71.49%	66.80%
Wikipedia-MF	63.26% (+0.93%)	69.01% (–3.5%)	66.01% (–1.2%)
Wikipedia-MD	71.43% (+13.96%)	76.45% (+6.94%)	73.85% (+10.55%)
Freebase	69.32%	75.62%	72.33%
Freebase-MF	69.47% (+0.22%)	75.21% (–0.54%)	72.22% (–0.15%)
Freebase-MD	69.26% (–0.09%)	77.27% (+2.2%)	73.05% (+1%)

(注：Wikipedia-MF表示在该实验中仅利用Wikipedia指称–实体字典进行指称识别但同时利用Wikipedia和Freebase实体特征进行实体消歧，以验证多特征对实体链接效果的影响，括号中的值是相对基于Wikipedia的实体链接效果的提升值；Wikipedia-MD表示在该实验中利用Wikipedia和Freebase指称–实体字典进行指称识别但仅利用Wikipedia实体特征进行实体消歧，以验证多指称–实体字典对实体链接效果的影响，括号中的值是相对基于Wikipedia的实体链接效果的提升值；Freebase-MF、Freebase-MD类似。)

从表2可以看出，利用多个指称–实体字典进行指称识别是实体链接效果提升的主要原因，其中召回率分别提升6.94%和2.2%，F值分别提升10.55%和1%；多特征仅对实体链接的准确率有较小提升(分别提升0.93%和0.22%)，对实体链接效果总体没有提升作用。

笔者分析了在基于Wikipedia+Freebase的实体链接实验中识别错误的查询主题，发现以下问题：

(1) 识别候选指称时可能过滤掉正确的指称。例如对于查询主题old coins，其正确的指称是coins，但是由于“old coins”覆盖度更大，并且符合最长匹配的原则，因此被保留。笔者发现“old coins”在Wikipedia中仅有一次作为指称出现，但是由于缺少相应的策略，造成指称识别错误。同因，espn sports、diabetes education也发生了指称识别错误。

(2) 获取候选实体时指称所指向的实体可能没被获取。例如对于查询主题website design hosting，在获取指称hosting的候选实体时，根据最大先验概率选取Wikipedia中该指称可能指向的前k个实体，然而这k个实体不包含指称hosting所指向的实体，因此造成候选实体选取错误，并且笔者发现Freebase中指称hosting可能指向的前k个实体包含该指称所指向的实体。同因，lymphoma in dogs、fact on uranus也发生候选实体选取错误。

(3) 未能正确地对实体进行消歧。例如对于查询主题obama family tree，基于Wikipedia+Freebase的实体链接未能正确地对其进行消歧，造成实体识别错误。笔者认为向消歧框架中加入更多特征或许能解决这类问题。

6　结　语

本文提出一种基于多知识库的实体链接方法，以Wikipedia和Freebase为例的实验结果表明，基于Wikipedia+Freebase的实体链接效果高于基于Wikipedia或Freebase的实体链接效果。本文存在两点不足，即基于词性过滤n-gram缺乏理论依据、数据集FACC1具有高准确率和低召回率的特点[8]。另外，本文的方法也可适用于其他知识库，例如对于YAGO，利用“HasWikipediaURL”关系可以构建YAGO实体与Wikipedia实体的映射字典，再结合笔者构建的Wikipedia实体与Freebase实体映射字典可构建三个知识库实体的映射字典；利用“means”关系收集YAGO实体的指称[6]并利用Wikipedia实体主页的锚文本统计指称指向其候选实体的次数[23]，从而构建相同结构的指称–实体字典；利用文本相关度、类型相关度(根据type关系和subClass关系计算的实体类型距离[23])等进行实体消歧。基于本文的结论，笔者认为基于Wikipedia+Freebase+YAGO的实体链接效果将会高于基于Wikipedia或Freebase或YAGO的实体链接效果。未来笔者将会探索更好的信息融合方式，以期进一步提升基于多知识库的实体链接效果。

参考文献：

[1] Zhang W, Sim Y C, Su J, et al. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling [C]. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain. 2011: 1909-1914.

[2] Pantel P, Fuxman A. Jigs and Lures: Associating Web Queries with Structured Entities [C]. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland,

Oregon, USA. 2011: 83-92.

[3] Lin T, Etzioni O. Entity Linking at Web Scale [C]. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, Montreal, Canada. 2012: 84-88.

[4] Welty C, Murdock J W, Kalyanpur A, et al. A Comparison of Hard Filters and Soft Evidence for Answer Typing in Watson [C]. In: Proceedings of the 11th International Conference on the Semantic Web. Springer-Verlag, 2012: 243-256.

[5] Bollacker K, Evans C, Paritosh P, et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge [C]. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, 2008: 1247-1250.

[6] Suchanek F M, Kasneci G, Weikum G. YAGO: A Core of Semantic Knowledge [C]. In: Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 697-706.

[7] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data [C]. In: Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Busan, Korea. 2007: 722-735.

[8] ClueWeb09 Related Data: Freebase Annotations of the ClueWeb Corpora, v1 (FACC1) [EB/OL]. (2013-11-04). [2015-11-24]. http://lemurproject.org/clueweb09/FACC1/.

[9] Brandão W C, Santos R L T, Ziviani N, et al. Learning to Expand Queries Using Entities [J]. Journal of the Association for Information Science and Technology, 2014, 65(9): 1870-1883.

[10] 陆伟, 武川. 实体链接研究综述[J]. 情报学报, 2015, 34(1): 105-112. (Lu Wei, Wu Chuan. Literature Review on Entity Linking [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 105-112.)

[11] Cucerzan S. Large-scale Named Entity Disambiguation Based on Wikipedia Data [C]. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007: 708-716.

[12] Milne D, Witten I H. Learning to Link with Wikipedia [C]. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM, 2008: 509-518.

[13] Ferragina P, Scaiella U. Tagme: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities) [C]. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Ontario, Canada. 2010: 1625-1628.

[14] Meij E, Weerkamp W, De Rijke M. Adding Semantics to Microblog Posts [C]. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining. ACM, 2012: 563-572.

[15] Sil A, Yates A. Re-ranking for Joint Named-entity Recognition and Linking [C]. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013: 2369-2374.

[16] Mihalcea R, Csomai A. Wikify!: Linking Documents to Encyclopedic Knowledge [C]. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, Lisboa, Portugal. 2007: 233-242.

[17] Zhang W, Su J, Tan C L, et al. Entity Linking Leveraging: Automatically Generated Annotation [C]. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, Beijing, China. 2010: 1290-1298.

[18] Pilz A, Paaß G. From Names to Entities Using Thematic Context Distance [C]. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, Scotland, UK. 2011: 857-866.

[19] Zheng Z, Li F, Huang M, et al. Learning to Link Entities with Knowledge Base [C]. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 483-491.

[20] Ratinov L, Roth D, Downey D, et al. Local and Global Algorithms for Disambiguation to Wikipedia [C]. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: 1375-1384.

[21] Shen W, Wang J, Luo P, et al. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge [C]. In: Proceedings of the 21st International Conference on World Wide Web, Lyon, France. 2012: 449-458.

[22] Han X, Sun L, Zhao J. Collective Entity Linking in Web Text: A Graph-based Method [C]. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China. 2011: 765-774.

[23] Hoffart J, Yosef M A, Bordino I, et al. Robust Disambiguation of Named Entities in Text [C]. In: Proceedingsof the

Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 782-792.

[24] Hachey B, Radford W, Curran J. Graph-Based Named Entity Linking with Wikipedia [C]. In: Proceedings of the 12th International Conference on Web Information System Engineering. 2011: 213-226.

[25] Guo Y, Che W, Liu T, et al. A Graph-based Method for Entity Linking [C]. In: Proceedings of the 5th International Joint Conferenceon Natural Language Processing, Chiang Mai, Thailand. 2011: 1010-1018.

[26] Gottipati S, Jiang J. Linking Entities to a Knowledge Base with Query Expansion [C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 804-813.

[27] Zhang W, Sim Y C, Su J, et al. NUS-I2R: Learning a Combined System for Entity Linking [C]. In: Proceedings of Text Analysis Conference 2010 Workshop, Gaithersburg, Maryland, USA. 2010.

[28] Chen Z, Ji H. Collaborative Ranking: A Case Study on Entity Linking [C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Scotland, UK. 2011: 771-781.

[29] Liu X, Li Y, Wu H, et al. Entity Linking for Tweets [C]. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2013.

[30] Wu C, Lu W, Zhou P. An Optimization Framework for Entity Recognition and Disambiguation [C]. In: Proceedings of the 1st International Workshop on Entity Recognition & Disambiguation. ACM, 2014: 105-110.

[31] Bunescu R C, Pasca M. Using Encyclopedic Knowledge for Named Entity Disambiguation [C]. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. 2006: 9-16.

作者贡献声明:

周鹏程: 方案设计, 实验实施, 论文起草以及修订;
武川: 方案设计和修订, 论文多次修订;
陆伟: 指导方案设计和论文写作, 论文多版本及最终版修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 周鹏程, 武川, 陆伟. 标注数据.xml. 主题的实体标注数据.

[2] 周鹏程, 武川, 陆伟. freebase 识别结果.txt. 基于 Freebase 的实体识别结果.

[3] 周鹏程, 武川, 陆伟. wikipedia 识别结果.txt. 基于 Wikipedia 的实体识别结果.

[4] 周鹏程, 武川, 陆伟. wikipedia_freebase 识别结果.txt. 基于 Wikipedia+Freebase 的实体识别结果.

[5] 周鹏程, 武川, 陆伟. freebase 指称–实体字典.txt. 基于 Freebase 构建的指称–实体字典.

[6] 周鹏程, 武川, 陆伟. wikipedia 指称–实体字典.txt. 基于 Wikipedia 构建的指称–实体字典.

[7] 周鹏程, 武川, 陆伟. freebase 实体-wikipedia 实体映射字典.txt. Freebase 实体与 Wikipedia 实体映射字典.

收稿日期: 2016-01-13
收修改稿日期: 2016-03-20

Entity Linking Method for Short Texts with Multi-Knowledge Bases: Case Study of Wikipedia and Freebase

Zhou Pengcheng[1] Wu Chuan[1] Lu Wei[1, 2]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper proposes an entity linking method using multi-knowledge bases, aiming at solving the problem of low coverage caused by entity linking with single knowledge base. [**Methods**] First, we generated n-gram of input text and obtained candidate mentions using part of speech and multi-mention-entity dictionary. Second, we generated and retained mention combinations of highest coverage which are not contained by other mention combinations. Third, we generated entity sequences and calculated their relevence degree using information from multi-knowledge bases. We listed entity sequence with the highest relevence degree as the final result. [**Results**] This case study showed that the Precision, Recall, and F-value of the entity linking based on Wikipedia+Freebase reaches 71.81%, 76.86%, and 74.25% respectively. [**Limitations**] Filtering n-gram based on part of speech lacked theoretical foundation, and the FACC1 dataset featured high precision but low recall. [**Conclusions**] Utilizing entity information from multi-knowledge bases can improve the performance of entity linking.

Keywords: Entity linking Knowledge base Wikipedia Freebase